Combining SVM Classifiers to Identify Investigator Name Zones in Biomedical Articles

Jongwoo Kim*, Daniel X. Le, and George R. Thoma
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA

ABSTRACT

This paper describes an automated system to label zones containing Investigator Names (IN) in biomedical articles, a key item in a MEDLINE® citation. The correct identification of these zones is necessary for the subsequent extraction of IN from these zones. A hierarchical classification model is proposed using two Support Vector Machine (SVM) classifiers. The first classifier is used to identify an IN zone with highest confidence, and the other classifier identifies the remaining IN zones. Eight sets of word lists are collected to train and test the classifiers, each set containing collections of words ranging from 100 to 1,200. Experiments based on a test set of 105 journal articles show a Precision of 0.88, 0.97 Recall, 0.92 F-Measure, and 0.99 Accuracy.

Keywords: Investigator Names, MEDLINE, Support Vector Machine, labeling, text classification, bibliographic information.

1. INTRODUCTION

The U.S. National Library of Medicine (NLM) maintains MEDLINE, a heavily used bibliographic database of 20 million citations to the biomedical journal literature. Each citation consists of bibliographic information such as article title, author names, affiliations, the abstract, grant numbers, databank accession numbers, etc. While NLM receives most such citations in XML format directly from journal publishers, key bibliographic information is often missing, requiring manual entry. We have therefore developed a system called Publisher Data Review System (PDRS) to automatically extract missing bibliographic information such as grant numbers, grant support information, databank accession numbers, Investigator Names (IN), etc. [1, 2, 3, 4].

In a typical biomedical article, the author zone is located close to the article title, and contains names of the authors. However, due to the increasingly collaborative nature of biomedical research, many investigators from various groups/organizations may collaborate in conducting the research. These group/organization names might also appear in the author or title zones, and the investigators affiliated with them would have their names listed somewhere else within the article, most likely toward the end, close to the references. In articles that list investigators, the number of investigator names average about 40, but may number several hundred.

The manual process to enter this data is time-consuming and error-prone. We have therefore proposed a system for automatic extraction of IN consisting of three steps. The system first divides an article into zones, next identifies zones containing IN, and then extracts IN from these zones. In this paper, we will focus on the second step, that of identifying (labeling) IN zones.

Algorithms commonly used for document labeling or named entity recognition include: Naïve Bayes algorithm [5] used for spam emails [6] and Web document classification [7], and Support Vector Machine (SVM) [8] used to categorize newswire documents, Medical Subject Headings (MeSH) [9], Web documents [10], and Reuters-21578 collection [11]. Since there are various types of zones with IN, supervised learning algorithms are proper in this case. We, therefore, use SVM classifiers in this work.

This paper is organized as follows. The definitions are given in Section 2. The details of our method using SVM classifiers are presented in Section 3. Performance evaluation measures are shown in Section 4. We report experimental results in Section 5, and conclusions in Section 6.

*jongkim@mail.nih.gov; phone 1 301 435-3227; fax 1 301 402-0341:

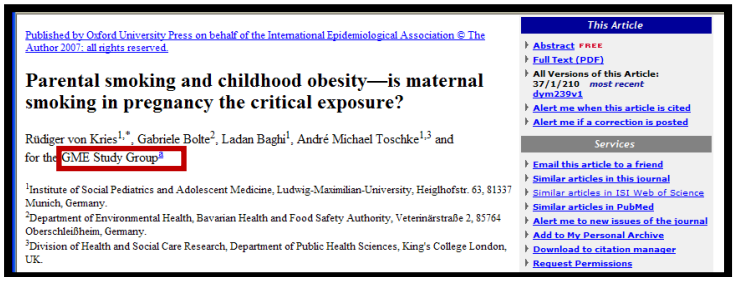
Document Recognition and Retrieval XIX, edited by Christian Viard-Gaudin, Richard Zanibbi,
Proc. of SPIE Vol. 8297, 829704 · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.910517

2. INVESTIGATOR NAME ZONE

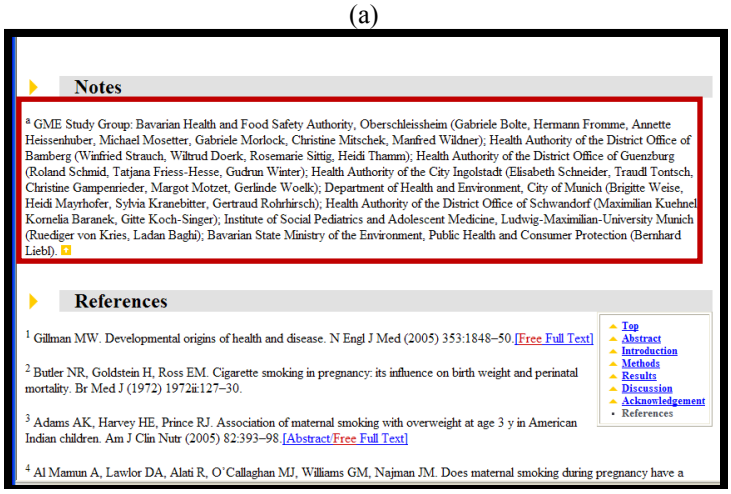
We define the group/organization names that appear in an author zone or in the title of an article as "Corporate Author". Investigators who are affiliated with the "Corporate Author" would have their names included in a zone (different from author zone) that is usually near the reference section or an affiliation zone in the article. In this paper, we refer to names of investigators as "Investigator Names (IN)" and zones containing these investigator names as "IN zones." Zones that do not contain IN will be considered "Non-IN zones." Sometimes, an author name can appear in both the author and IN zones if the author is affiliated with a Corporate Author.

Figure 1 shows an example of "Corporate Author" and its corresponding "IN zone." Figure 1(a) shows the Corporate Author "GME Study Group" located in the author zone, and Figure 1(b) shows the corresponding "IN zone" located right above the reference section. Figure 2 shows examples of "IN zones" in red boxes and "Non-IN zones" in green boxes. Figure 2(a) shows two "IN zones" and Figure 2(b) shows eight "IN zones" followed by four "Non-IN zones." Of these four zones, two (green boxes) are similar to "IN zones" and the others which are not boxed ("Liasions" and "Centers for Disease …") are not.

In general, Investigator Names are grouped together in a single "IN zone" as shown in Figure 1(b); however they might also appear in multiple "IN zones" as shown in Figures 2(a) and 2(b). The information in "IN zones" does not follow any particular format. Some "IN zones" might contain only investigator names, but others might have investigator names in addition to their affiliations, and other information. Moreover, the text size of each "IN zone" depends on the total number of Investigator Names. As seen in Figure 1(b), the contents of the "IN zone" including investigator names and their affiliations are similar to those of author zones. In Figure 2(b), each "IN zone" consists of only one investigator name and its format is similar to that of names in the reference section. As a result, information from neighboring zones, including "Corporate Author" information, section names such as abstract, introduction, references, etc. and zone locations within an article, are helpful in identifying "IN zones" correctly.

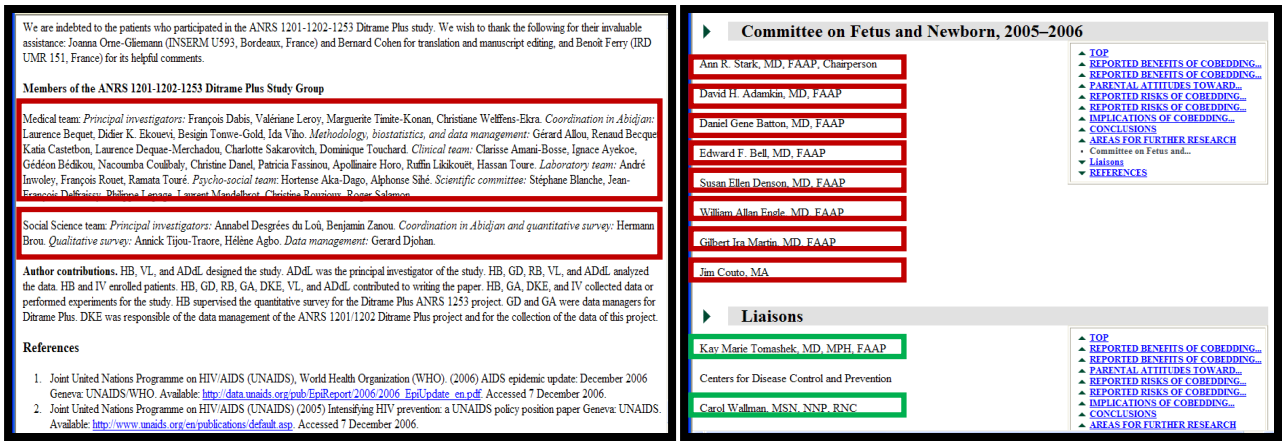


(a)



(b)

Figure 1: (a) An author zone with the Corporate Author "GME Study Group." (b) The corresponding "IN zone" is located right above the reference section.



(a) (b)

Figure 2: Examples of articles with multiple "IN zones" and "Non-IN zones." (a) Two "IN zones" (red boxes) next to each other. (b) Eight "IN zones" (red boxes) next to each other followed by four "Non-IN zones": Two "Non-IN zones" (green boxes) are similar to "IN Zones" and the others ("Liaisons" and "Centers for Disease …") are not.

3. OUR APPROACH

3.1 Module to label Investigator Name Zones

The proposed module shown in Figure 3 consists of two classifiers: IN-Classifier and Post-Classifier. The IN-Classifier labels zones based on zone features and is designed to classify a single "IN zone" similar to the one shown in Figure 1. The Post-Classifier uses the IN-Classifier's results in addition to information from neighboring zones to obtain the final labeling results and is designed for classifying multiple "IN zones" similar to those shown in Figure 2. The main features of an "IN zone" are the names of investigators so the number of names identified in a zone plays an important role in recognizing a zone as an "IN zone." When an "IN zone" is divided into multiple zones, it is harder to label them. As a result, the Post-Classifier has to exploit information from neighboring zones to improve the classification rate for multiple "IN zones."

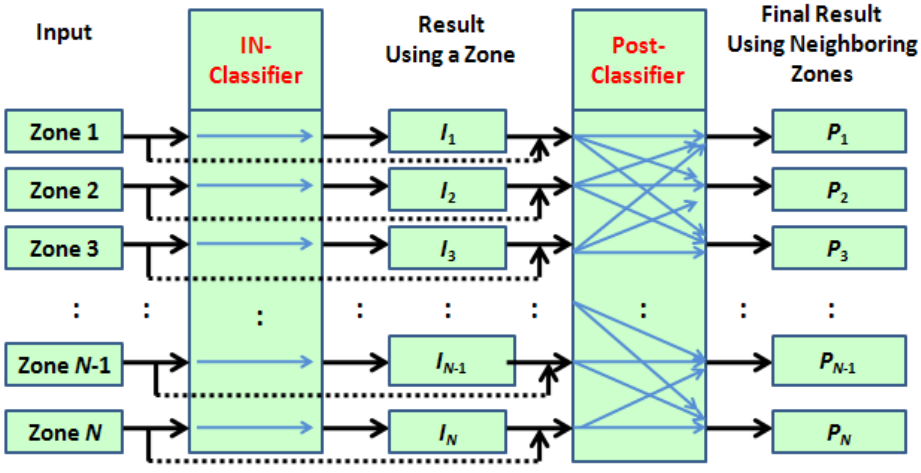


Figure 3. The proposed IN zone labeling module. I_i is the result of the IN-Classifier for zone *i,* and P_i is the result of the Post-Classifier for zone *i*.

The module workflow is as follows.

Let I_i be the estimation result ($0 \le I_i \le 1$) of the IN-Classifier for zone i (Z_i) and
P_i be the estimation result ($0 \le P_i \le 1$) of the Post-Classifier for zone i (Z_i) in an article, where $i = 1, 2, \ldots, n$.
First, estimate all I_i in an article, where $i = 1, 2, \ldots, n$.
Second, estimate $m = \arg\max_{i \in \{1,2,\ldots,n\}} I_i$
Third, set $h = m$, $k = m$, and threshold t ($t = 0.5$ in our case)
Fourth, set $h = h - 1$.
    If ($P_h \ge t$),
        Label Z_h as an "IN zone" and continue the fourth step.
    Else
        Stop labeling the next zone.
Fifth, set $k = k + 1$.
    If ($P_k \ge t$),
        Label Z_k as an "IN zone" and continue the fifth step.
    Else
        Stop labeling the next zone.

3.2 Feature extraction

Several categories of descriptive words may be used as features. IN zones are usually composed of author names and affiliation words and contain common words from lexicons that may appear in Non-IN zones. Therefore, we create eight word lists from MEDLINE as shown in Table 1. From these word lists, we extract fourteen features from each zone. Table 2 shows some of these features.

Table 1. Word lists used for labeling an "IN zone."

Word Lists	**Explanation/Examples**
Common words in Corporate Author	Committee, Investigator, Group, etc.
Section name	Acknowledgment, Notes, etc.
Last name	Arison, Barret, Chay, Digman, Elgar, Forbes, Grossmann, etc.
First name	Adam, Bent, Carol, Dave, Ema, Frank, Gary, etc.
Title	M.D., R.N., Ph.D., M.S., etc.
Affiliation	Department, School, Hospital, etc.
Grant Support	Grant, Support, Fund, etc.
Other words	Activation, broadly, case, delivered, project, separation, etc.

Table 2. Features used for labeling an "IN zone."

Features	**Explanation/Examples**
Corporate Author name	Complete name
Common Corporate Author word	Committee, Investigator, Group, etc.
Section Name	Acknowledgment, Notes, etc. (excluding References)
References (Section Name)	The word "References" as title of that section.
Zone having more than two words	
Frequency of Last and First Name	Proportion of these in a zone.
Frequency of All Names	Proportion of these in a zone.
Frequency of Title	Proportion of these in a zone.
Frequency of Affiliation	Proportion of these in a zone.
Frequency of punctuation mark	Proportion of these in a zone.
Frequency of total words matched	Proportion of these in a zone.
Frequency of Grant Support	Proportion of these in a zone.
Normalized Frequency of All Names	
Normalized Frequency of Number of Words	

3.3 SVM classifiers

We use the LIBSVM [12, 13] library for SVM classifiers and radial basis functions (RBF) as their kernel functions. In the case of parameters (C, γ), the library automatically sets its own parameters after the optimization process.

Given training vectors $\boldsymbol{x}_i \in R^n$, $i = 1, 2,\ldots,l$, in two classes $y_i \in \{1,-1\}$ (1 means relevant class and -1 means non-relevant class), the SVM tries to solve the following problem.

$$\min_{\boldsymbol{W},b,\xi} \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} + C\sum_{i=1}^{l}\xi_i \qquad \text{subject to } y_i(\boldsymbol{W}^T\phi(\boldsymbol{x}_i) + b) \geq 1-\xi_i,\ \xi_i \geq 0,\ i = 1, 2,\ldots, l. \tag{1}$$

In the equation, $\boldsymbol{W}$ is a normal (weight) vector, ξ_i is a penalty parameter of the error term, and b is the bias.

When $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$ is the kernel and $\phi(\boldsymbol{x}_i)$ is a function mapping $\boldsymbol{x}_i$ into a higher dimensional space, the decision function is

$$\operatorname{sgn}\left(\sum_{i=1}^{l} y_i\alpha_i K(\boldsymbol{x}_{i,}\boldsymbol{x}) + b\right), \text{ where } \alpha_i \text{ is a constraint and } 0 \leq \alpha_i \leq 1. \tag{2}$$

We use the following sigmoid function to convert a result obtained in (2) into a value between 0 and 1.

$$G_{SVM}(\mathbf{x}) = \frac{1}{1+e^{-\alpha t}}, \text{ where } t = \sum_{i=1}^{l} y_i\alpha_i K(\boldsymbol{x}_{i,}\boldsymbol{x}) + b \tag{3}$$

4. PERFORMANCE EVALUATION MEASURES

We use four measures, Precision, Recall, F-Measure, and Accuracy, to evaluate the performance of SVM classifiers and the merging operators. The measures are expressed as follows:

Precision = $TP/(TP+FP)$,
Recall = $TP/(TP+FN)$,
F-Measure = $2\times Precision\times Recall/(Precision+Recall)$,
Accuracy = $(TP+TN)/(TP+TN+FP+FN)$,`

where *TP*, *TN*, *FP* and *FN* stand for the numbers of "true-positives", "true-negatives", "false-positives", and "false-negatives", respectively.

5. RESULTS AND DISCUSSION

We collect two sets of training data: one for the IN-Classifier and the other for the Post-Classifier. We combine all "IN zones" in an article together into one zone to serve as the training set for the IN-Classifier. However, we use multiple "IN zones" as well as the combined zone to train the Post-Classifier. For example, for training, the IN-Classifier uses the "IN zone" resulting from combining the eight separate "IN zones" shown in Figure 2(b), and the Post-Classifier uses all eight individual "IN zones" as well as the combined "IN zone". To train "Non-IN zones" for both classifiers, we randomly select zones from a set of training articles. Here, two neighboring zones (one zone before and one after an IN zone) are used to train the Post-Classifier.

There is relatively little training data since not many articles have Investigator Names (IN) and also because IN has only recently been required, so that there are few MEDLINE citations containing IN. As a result we have only 159 articles for training. 159 "IN zones" and 433 "Non-IN zones" are used to train the IN-Classifier and 448 "IN zones" and 530 "Non-IN zones" used to train the Post-Classifier. Table 3 shows training results for the two classifiers.

To test the two SVM classifiers and the proposed module, we collect 105 journal articles containing 147 "IN zones" and 22,095 "Non-IN zones." Tables 4 and 5 show the test results.

For the IN-classifier, Precision is low, but Recall is better (0.88), and Accuracy is high (0.99). Since the IN-Classifier assumes only one "IN zone" in an article, it does not work well for any article having multiple "IN zones" and it often misclassifies "IN zones" as "Non-IN zones." The IN-Classifier also shows several over-labeling problems, that is, it misclassifies some names appearing in author and reference zones as Investigator Names.

The Post-Classifier trained to label multiple "IN zones" yields a high Recall rate, but Precision is lower than that of the IN-Classifier. This is reasonable because multiple "IN zones" are used for training and features in these zones are similar to those of author zones and reference zones. However, the Post-Classifier also yields a very high Accuracy rate (0.98).

The last columns of Tables 4 and 5 show the performance of the proposed module achieved by combining the two SVM classifiers. The Precision rate improves to 0.88; Recall rate 0.97; F-Measure rate 0.92; and Accuracy rate 0.99.

Notwithstanding this good performance, the proposed module does produce some false positive and false negative errors. False-negative errors generated by the IN-Classifier cause false-positive errors in the Post-Classifier. In the case of five false-positive errors from five articles, the IN-Classifier estimates confidences of all "IN zones" above a threshold to be labeled as "IN zones," but each of these "IN zones" does not have the highest confidence in each article, but it assigns highest confidence to two author zones, one affiliation zone, one reference zone, and one other zone. Since the IN-Classifier only considers a zone with a highest confidence as an "IN zone" for the Post-Classifier, all the real "IN zones" are not labeled as such. In the case of false-positive errors, some are caused by false-negative errors from the IN-Classifier and some are errors made by the Post-Classifier. We can demonstrate the second case by using Figure 2(b). This figure shows eight multiple "IN zones" (red boxes) and four "Non-IN zones." Of the four zones, two (green boxes) are similar to "IN zones" and the others (no box) are not. The Post-Classifier labels them all as "IN zones." The zones "Liasions" and "Centers for Disease Control and Prevention" are also labeled as "IN zones" since information from two neighboring zones strongly suggests "IN zones."

In order to correct this type of error, we will build heuristic rules based on information from neighboring zones such as keywords (e.g., subtitles, references, etc.), zone locations, etc. Figure 4 shows an example of a false-positive and false-negative error. Figure 4(a) shows an "IN zone" (red box) misclassified as a "Non-IN zones" and Figure 4(b) shows a reference zone (green box) classified as an "IN zone." The IN-Classifier estimates these two zones as candidates for "IN zones" with 0.84 confidence for an "IN zone" in Figure 4(a), and 0.94 confidence for a reference zone in Figure 4(b). However, the reference zone has higher confidence than the real "IN zone" because it contains more names than the other.

Table 3. Training results.

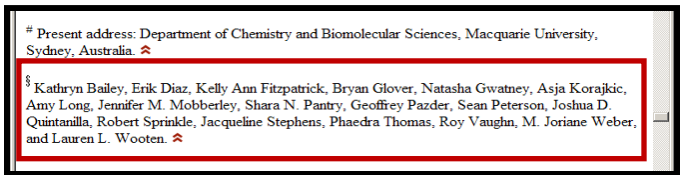
	IN-Classifier		**Post-Classifier**
IN Zone	159		448
Non-IN Zone	433		530
Precision	0.99		0.99
Recall	0.97		1.00
F-Measure	0.98		0.99
Accuracy	0.99		0.99

Table 4. Performance of the individual classifiers and the proposed module. (105 articles used.)

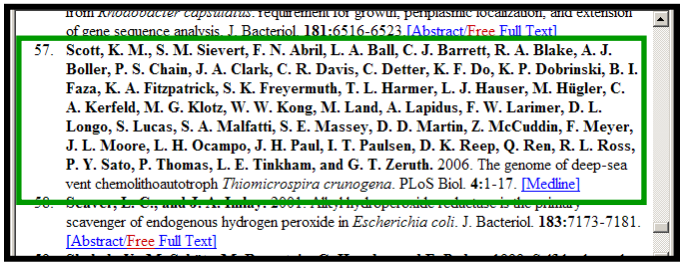
	IN-Classifier			**Post-Classifier**			**Proposed Module**	
	True	**False**		**True**	**False**		**True**	**False**
IN Zone (147)	130	17		145	2		141	5
Non-IN Zone (22,095)	141	21,954		348	21,747		19	22,076

Table 5. Performance measures for the individual classifiers and the proposed module.

	IN-Classifier		**Post-Classifier**		**Proposed Module**
Precision	0.48		0.29		0.88
Recall	0.88		0.99		0.97
F-Measure	0.62		0.45		0.92
Accuracy	0.99		0.98		0.99



(a)



(b)

Figure 4. An example of false-positive and false-negative errors in an article. (a) An "IN zone" (red box) is misclassified as a "Non-IN zone." (b) A reference zone (green box) is misclassified as an "IN zone."

6. CONCLUSIONS

In this paper, we describe a hierarchical labeling module consisting of two SVM classifiers to automatically classify zones that contain Investigator Names in an online biomedical article, as a preliminary step to extracting Investigator Names to be included in a MEDLINE citation.

We collect fourteen word features to train and test the two SVM classifiers from words that occur most frequently in "IN zones" and in "Non-IN zones." The IN-Classifier is used to label a zone with highest confidence as an IN zone and the Post-Classifier is used to label multiple IN zones using neighboring zones' information and results from the IN-Classifier. The proposed module combining the two SVM classifiers shows relatively good performance. Precision is somewhat low (0.88), but, Recall and F-Measure are relatively high (0.97 and 0.92). The accuracy is also high (0.99). The IN-Classifier creates false-negative errors by selecting only one zone as an IN zone candidate in an article, and some false-positive errors result from references in the article. For the Post-Classifier, false-positive errors are generated by multiple IN zones.

A future task therefore is to label more candidates as IN zones while applying heuristic rules using Corporate Author information, section names, keywords, etc. to resolve the problem of false-negative errors due to multiple IN zones. We also plan to use more features such as journal names, years, pagination, etc. to eliminate false-positive errors generated by reference sections.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

REFERENCES

[1] Kim, J., Le, D. X., and Thoma, G. R., "Naïve Bayes and SVM Classifiers for Classifying Databank Accession Number Sentences from Online Biomedical Articles," IS&T/SPIE's 22nd Annual Symposium on Electronic Imaging, San Jose, CA, 7534: 75340U-1-8 (2010).

[2] Kim, J., Le, D. X., and Thoma, G. R., "Inferring Grant Support Types From Online Biomedical Articles," Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems, Albuquerque, New Mexico. (2009).

[3] Kim, J., Le, D. X., and Thoma, G. R., "Naïve Bayes Classifier For Extracting Bibliographic Information From Biomedical Online Articles," Proc. International Conference on Data Mining, Las Vegas, Nevada, USA, II: 373-8 (2008).
[4] "Technical Memorandum 484: Investigator Names," National Institutes of Health, National Library of Medicine, (2008).
[5] Lewis, D. D., "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval," *ECML*, The Tenth European Conference on Machine Learning, 4-15 (1998).
[6] Madigan, D., "Statistics and the war on spam," Statistics: A Guide to the Unknown, 4th Ed. (R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern, eds.), Thomson Brooks/Cole, Belmont, CA, 135–147 (2005).
[7] McCallum, A. and Nigam, K., "A Comparison of Event Models for Naïve Bayes Text Classification," Proc. the AAAI-98 Workshop on Learning for Text Categorization, 577 (1998).
[8] Burges, C. J. C., "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, 121-167 (1998).
[9] Joschims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE, 137-142 (1998).
[10] Gabrilovich, E. and Markovitch, S., "Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5", ICML'04, 321- 328 (2004).
[11] Dumais, S., Platt, J., Heckerman, D. and Sahami, M., "Inductive learning algorithms and representations for text categorization," Proc. CIKM-98, 7th ACM International Conference on Information and Knowledge Management, Washington, 148–155 (1998).
[12] Chang, C. C. and Lin, C. J., "LIBSVM: a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, (2001).
[13] Johnson, M., "SVM.NET", Software available at http://www.matthewajohnson.org/index.html, (2008).